DOCUMENT RESUME

ED 393 911 TM 024 912

AUTHOR Rudas, Tamas; Zwick, Rebecca
TITLE Estimating the Importance of Differential Item Functioning. Program Statistics Research Technical Report No. 95-3.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-95-33
PUB DATE Oct 95
NOTE 32p.; Supported in part by a grant from the Hungarian National Science Foundation.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Estimation (Mathematics); Goodness of Fit; *Item Bias; *Maximum Likelihood Statistics; Physics; Simulation; Test Construction; *Test Items
IDENTIFIERS Advanced Placement Examinations (CEEB); *Contingency Tables; Item Bias Detection

ABSTRACT

A method is proposed to assess the importance of differential item functioning (DIF) by estimating the largest possible fraction of the population in which DIF does not occur, or equivalently, the smallest possible portion of the population in which DIF may occur. The approach is based on latent class (C. C. Clogg, 1981) or mixture concepts, and was proposed by T. Rudas, C. C. Clogg, and B. G. Lindsay (1994) in the context of assessing the fit of an arbitrary model to a contingency table. Application of this procedure produces an estimate of the minimum proportion of the population that would have to be removed to make the rest of the population free from DIF, as well as information about the portion of the population that is the source of DIF. Simple methods for maximum likelihood estimation are described. Numerical results are presented for a simulated data set and actual data from the 1993 Advanced Placement Physics examination. (Contains 3 tables and 27 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

RR-95-33

ED 393 911

Estimating The Importance of Differential Item Functioning

Tamás Rudas
Eötvös University and TÁRKI, Budapest

Rebecca Zwick
Educational Testing Service

PROGRAM
STATISTICS
RESEARCH

Technical Report No. 95-3

Educational Testing Service
Princeton, New Jersey 08541

TM024912

BEST COPY AVAILABLE

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

Estimating The Importance of Differential Item Functioning

Tamás Rudas
Eötvös University and TÁRKI, Budapest

Rebecca Zwick
Educational Testing Service

Program Statistics Research
Technical Report No. 95-3

Research Report No. 95-33

Educational Testing Service
Princeton, New Jersey 08541

October 1995

Copyright © 1995 by Educational Testing Service. All rights reserved.

Estimating the importance of differential item functioning

Tamás Rudas*
Eötvös University and TÁRKI, Budapest

Rebecca Zwick
Educational Testing Service, Princeton

*This research was initiated when Rudas was a Visiting Scholar in the Research Statistics Group of the Educational Testing Service. Rudas's work was also supported in part by Grant No. OTKA T-016032 from the Hungarian National Science Foundation.

Estimating the importance of differential item functioning

Several methods have been proposed to detect differential item functioning (DIF), an item response pattern in which members of different demographic groups have different conditional probabilities of answering a test item correctly, given the same level of ability. In this paper, the mixture index of fit, proposed by Rudas, Clogg, and Lindsay (1994) is used to estimate the fraction of the population for which DIF occurs, and this approach is compared to the Mantel-Haenszel (1959) test of DIF developed by Holland (1985; see Holland & Thayer, 1988). The proposed estimation procedure, which is noniterative, can provide information about which portions of the item response data appear to be contributing to DIF.

Key words: differential item functioning, Mantel-Haenszel test, maximum likelihood estimation, mixture index of fit

Acknowledgement: The authors are indebted to Nicholas T. Longford, Ming-mei Wang, two anonymous referees and the Associate Editor for comments on an earlier version of this manuscript, to Dorothy T. Thayer for computing summary statistics for the examples and to the College Board's Advanced Placement Program® at ETS for providing data from the AP Physics B examination.

Introduction

The absence of differential item functioning (DIF) is regarded as an important aspect of test fairness by most educational researchers. The extensive literature on the detection and measurement of DIF is reviewed in Holland and Wainer (1993) and Camilli and Shepard (1994).

In this paper we propose to assess the importance of differential item functioning by estimating the largest possible fraction of the population in which DIF does not occur, or, equivalently, the smallest possible portion of the population in which DIF may occur. This approach is based on latent class (see Clogg, 1981) or mixture concepts and was proposed by Rudas, Clogg, and Lindsay (1994) in the general context of assessing the fit of an arbitrary model to a contingency table.

Let H be any model or hypothesis for a contingency table. Then any distribution P can be represented as

$$P = (1-\pi)\Phi + \pi\Psi \,, \tag{1}$$

where Φ is a distribution in H, Ψ is an arbitrary distribution, and $0 \le \pi \le 1$. The above representation is not unique. The mixture index of fit π^* is defined as the minimum possible value of π,

$$\pi^* = \inf \{ \pi : P=(1-\pi)\Phi+\pi\Psi, \ \Phi \in H \} \,,$$

and it is the smallest possible fraction of the population outside the model of interest, H. Rudas, Clogg, and Lindsay (1994) described a general method of obtaining maximum likelihood estimates of π^* and

of constructing confidence intervals. The nonrestricted distribution, Ψ, describes residuals, though not in the standard sense, and π^* is the total weight of these residuals. Ordinarily, residuals are defined with respect to a model that is assumed to hold **in the entire population**. By contrast, the residuals in this approach are defined in the context of representation (1), which is always true. The Ψ residuals describe the distribution in the part of the population in which hypothesis H is not true. Various interpretations of Ψ are discussed in Clogg, Rudas, and Xi (1995). In the present paper, the residuals will be used to identify parts of the population in which evidence of DIF exists.

An extension of the approach of Rudas, Clogg, and Lindsay (1994) will be used to compare the fits of nested models using a measure of the relative fit of a model against a restricted alternative (see also Clogg, Rudas, & Xi, 1995). This will be applied to the "no DIF" and "uniform DIF" (see Mellenbergh, 1982, Holland, 1985) hypotheses of the Mantel-Haenszel (MH) type.

Application of the procedure proposed in this paper produces an estimate of the minimum proportion of the population that would have to be removed in order to make the rest of the population free from DIF, as well as information about the specific portion of the population that is the apparent source of DIF in the above sense. This type of result may be more interpretable than conventional DIF statistics and may provide information that can be used to modify test items.

The paper is organized as follows: the next section formulates the hypotheses of no DIF and uniform DIF as MH-type hypotheses for a three-dimensional contingency table. Then simple methods for maximum

likelihood (ML) estimation of π^* under these hypotheses will be described, along with a method for testing the hypothesis that the fraction of the population that is free from DIF is greater than a specified value. The conclusions that can be drawn from inspecting the $\hat{\pi}^*$ values and $\hat{\Psi}$ residuals will also be discussed. The next section will present numerical results for two data sets -- a simulated data set and a set of examinee responses to the 1993 Advanced Placement Physics B Exam. The last section discusses relative advantages and disadvantages of using the mixture index of fit π^* in this context.

The hypothesis of no differential item functioning

Let A and B be two groups of respondents, often labeled as the focal and reference groups. The focal group is the group of primary interest and the reference group serves as a basis for comparison.

The analysis of DIF can be conducted by comparing the reference and focal group odds of answering the item correctly, conditional on a measure of ability, such as a test score. Under the hypothesis of no DIF, group membership and item response (correct or incorrect) are conditionally independent, given ability. The following table gives the notation for the conditional probabilities at level j of the matching test score

		Response	
		Correct	Incorrect
Group	A	P_{ACj}	P_{AIj}
	B	P_{BCj}	P_{BIj}

The hypothesis of no DIF is

$$(2) \qquad H \text{ (no DIF)} : \quad \frac{P_{ACj}P_{BIj}}{P_{AIj}P_{BCj}} = \alpha = 1 \text{ for } j=1,\ldots,J.$$

Holland (1985) suggested the use of the Mantel-Haenszel procedure for testing the hypothesis of no DIF (see also Holland, & Thayer, 1988). The Mantel and Haenszel (1959) chi-square test approximates the uniformly most powerful unbiased test of the null hypothesis against the alternative that the conditional odds ratios (see Rudas & Leimer, 1992) in (2) are all equal to a common value other than one (Holland, & Thayer, 1988), which is the hypothesis of uniform DIF:

$$(3) \qquad H \text{ (uniform DIF)} : \quad \frac{P_{ACj}P_{BIj}}{P_{AIj}P_{BCj}} = \alpha \ (\neq 1) \text{ for all } j.$$

The amount of DIF, as measured by the conditional odds ratio, is assumed to be constant over all levels of the matching variable.

When the sample size of the focal group is much smaller than the sample size of the reference group, the method for fitting the same log-linear model to two groups of very different sizes described in Rudas (1991) may be applied instead of testing (2) against (3).

Holland and Thayer (1988) discussed the relative advantages of testing (2) against (3) over other methods of testing for the presence of DIF (see also Zwick, 1990). They proposed the use of a transformation of the Mantel and Haenszel (1959) odds ratio estimator (i.e. the estimator under (3)), to measure the amount of DIF. In practice, a combination of the MH chi-square and odds ratio estimate is often used to assess the degree of DIF in an item (see

Zieky, 1993).

In the next section we provide an alternative way of assessing the amount of DIF by estimating the smallest fractions of the population that have the property that their complements can be described by hypotheses (2) and (3), respectively. The comparison of these two fractions can be used as a measure of the relative fits of hypotheses (2) and (3).

The hypotheses considered in this section can be extended to items with more than two possible scores, such as partial credit items or items that are scored on an ordinal scale. Within each level of the matching variable, the data can be represented as a 2xL table, where L is the number of options. In this case, the association structure can be described by considering either the conditional means of the two groups (e.g., see Zwick, Donoghue, & Grima, 1993a; Zwick & Thayer, in press) or the set of conditional odds ratios pertaining to these 2xL tables (Zwick, Donoghue, & Grima, 1993b). These can be the odds ratios based on neighboring columns (see Goodman, 1979) or on the reference cell approach (see Rudas, 1991). The methodology discussed in the next section can be applied in these cases as well, but iterative procedures are needed for fitting the models of no DIF or uniform DIF. A fourth, unobserved variable is introduced, showing whether or not an observation came from the part of the population in which the hypothesis holds. Then the EM algorithm (Dempster, Laird, & Rubin, 1977) can be applied to fit the mixture in (1) with various trial values of π. The value of $\hat{\pi}^*$ is the smallest value with which perfect fit can be achieved. This procedure is described in Rudas, Clogg, and Lindsay (1994) in a general form and will not be discussed here. On the other hand, when the responses are classified only as correct or incorrect, the ML estimate for π^* has

a closed form for the hypothesis of no DIF and can be obtained as the result of a finite-step maximization procedure for the hypothesis of uniform DIF. These procedures are considered in the next section.

Estimating the fraction of population outside of the hypotheses of no DIF and uniform DIF

The goal of the π^* approach, sketched briefly in the introduction, is to consider the observed table of frequencies and take away the smallest possible fraction of observations, so that what remains corresponds to the hypothesis of interest exactly. Note that the exact correspondence to the hypothesis which results does not imply that the procedure overfits the model; rather it is a consequence of the fact that representation (1) always holds true with an appropriate value of π. The ratio of the number of observations removed to the sample size is the ML estimate of the mixture index of fit π^* and the distribution of the portion of observations that was taken away is the ML estimate of Ψ, where Ψ is the distribution in that part of the population in which the hypothesis of interest does not hold (Rudas, Clogg, & Lindsay, 1994).

In the case of model (2), this leads to the following algorithm. For every level j of the matching variable, consider the table of observed frequencies and suppose that none of the entries are equal to zero. If the observed conditional odds ratio

$$\hat{\alpha}_j = \frac{f_{ACj} f_{BIj}}{f_{AIj} f_{BCj}}$$

is greater than $\alpha=1$, only the smaller of f_{ACj} and f_{BIj} needs to be

reduced. The smaller of these, $g_{smj}=\min(f_{ACj}, f_{BIj})$, must be reduced by

$$d_j = g_{smj}\,(1-\alpha/\hat{\alpha}_j) .$$

When $\hat{\alpha}_j$ is less than $\alpha=1$, only the smaller of f_{AIj} and f_{BCj} needs to be reduced. The smaller of these, $h_{smj}=\min(f_{AIj}, f_{BCj})$, must be reduced by

$$d_j = h_{smj}\,(1-\hat{\alpha}_j/\alpha) .$$

See Clogg, Rudas, and Xi (1994) for related discussion. The ML estimate of π^* for model (2) can be obtained as

$$\hat{\pi}^*(\text{no DIF}) = (1/N) \sum_j d_j ,$$

where N is the total sample size.

When $f_{ACj}=f_{BIj}$ or $f_{AIj}=f_{BCj}$, either one of the frequencies can be reduced. The cell of the conditional table in which the frequency is reduced is not uniquely defined, but the amount of decrease, and therefore the value of $\hat{\pi}^*$, are uniquely defined.

To design a simple algorithm yielding $\hat{\pi}^*$(uniform DIF), consider (3) as the union of infinitely many hypotheses:

$$H\,(\text{uniform DIF}) = \bigcup_{\substack{\alpha\in R \\ \alpha\neq 1}} H_\alpha = \bigcup_{\substack{\alpha\in R \\ \alpha\neq 1}} \{P:\ \frac{P_{ACj}P_{BIj}}{P_{AIj}P_{BCj}} = \alpha \} .$$

For $\alpha\neq\beta$, H_α and H_β are disjoint. Therefore, one may obtain $\hat{\pi}^*$ for (3) by first fixing α, finding $\hat{\pi}^*(H_\alpha)=\hat{\pi}^*(\alpha)$, and then taking the

infimum over the possible values of α. Note that H_α is a prescribed conditional interaction model in the three-way table (see Rudas, 1991).

For arbitrary but fixed α, the algorithm to find $\hat{\pi}^*(\alpha)$ is exactly like the one described above for hypothesis (2). This yields a $\hat{\pi}^*(\alpha)$ value and the ML estimate under hypothesis (3) can be obtained as

$$\hat{\pi}^*(\text{uniform DIF}) = \inf_{\alpha \neq 1} \hat{\pi}^*(\alpha) \; . \tag{4}$$

There is, however, no need to minimize over all positive $\alpha \neq 1$ values. It can be assumed without loss of generality that the ability levels are indexed by j in ascending order, that is $\hat{\alpha}_j \leq \hat{\alpha}_{j+1}$, for every j. If for some j, $\hat{\alpha}_j < \hat{\alpha}_{j+1}$ and $\hat{\alpha}_j \leq \alpha \leq \hat{\alpha}_{j+1}$, then

$$N\,\hat{\pi}^*(\alpha) = \sum_{i \leq j} h_{smi}\,(1-\hat{\alpha}_i/\alpha) + \sum_{i \geq j+1} g_{smi}\,(1-\alpha/\hat{\alpha}_i) \; .$$

In the range $\hat{\alpha}_j < \alpha < \hat{\alpha}_{j+1}$ the first derivative of the above function is positive, and the second derivative is negative, implying that in the range $\hat{\alpha}_j < \alpha < \hat{\alpha}_{j+1}$ the function $\hat{\pi}^*(\alpha)$ is convex. Therefore, in the range $\hat{\alpha}_j \leq \alpha \leq \hat{\alpha}_{j+1}$ the function $\hat{\pi}^*(\alpha)$ has its minimum either for $\alpha = \hat{\alpha}_j$ or for $\alpha = \hat{\alpha}_{j+1}$. Also, the minimum in (4) cannot occur for an α value outside of the range of the observed $\hat{\alpha}_j$ values, because for $\alpha < \alpha_1$, $\hat{\pi}^*(\alpha) > \hat{\pi}^*(\alpha_1)$, and for $\alpha > \alpha_J$, $\hat{\pi}^*(\alpha) > \hat{\pi}^*(\alpha_J)$. Therefore, it suffices to inspect only the values of $\hat{\pi}^*(\alpha)$ at the observed ability levels.

$$\hat{\pi}^*(\text{uniform DIF}) = \min_{\alpha = \hat{\alpha}_1, \ldots, \hat{\alpha}_J} \hat{\pi}^*(\alpha) \; .$$

Note that the estimates for π^* do not depend on the sample size as do the chi-square values for the hypotheses of independence or

conditional independence. If two samples have the same relative frequencies, the estimates of the mixture index of fit π^* are the same.

The above algorithms assume that there are no zero observed frequencies in the data. If, for a given level of the matching variable, zeros occur in both cells of the same column (i.e. either everybody in both groups, or nobody in either group could answer the item correctly), this can be regarded as inconsistent with DIF; these 2x2 tables may be omitted from the analysis. Two other ways to eliminate zero cells which may be appropriate in some instances are combining the data across two or more levels of the matching variables (Donoghue & Allen, 1993) or smoothing the data by using a suitable prior or by adding small constants to the empty cells (Agresti, 1990).

Having estimated π^*(no DIF) and π^*(uniform DIF), several inferential procedures are feasible. These parameters can be interpreted as the smallest possible fractions of the population that cannot be described by the model. The values of π^* can be used as measures of the misfit of the respective models, i.e. as measures of the amount of DIF. Also, these measures can be compared across items.

The pattern of the residual Ψ, i.e., the locations and relative sizes of the amounts that were removed from the conditional tables, provide information about where (in terms of ability level, group membership, and item response) DIF occurs.

If the hypothesis of uniform DIF is extended to include the case of α=1, then hypothesis (2) is nested in hypothesis (3) and $\hat{\pi}^*$(no DIF) $\geq$ $\hat{\pi}^*$(uniform DIF). The difference between these two values can be

used as a measure of how much better (3) fits the data than (2) does; i.e. what fraction of the population is lost by restricting the value of the common conditional odds ratio to one.

The above inferential procedures are illustrated in the next section.

In some cases, testing the hypothesis that the proportion of the population in which DIF is present is less than a specific value, say, $\bar{\pi}$ may be of interest. This can be done by fitting the model

$$P = (1-\bar{\pi})\Phi + \bar{\pi}\Psi, \quad \Phi \in H(\text{no DIF})$$

to the data. To fit this model, standard latent class techniques can be used, which involve defining a fourth, unobserved, variable that identifies whether an observation came from the distribution Φ or from the distribution Ψ, and applying the EM algorithm (Dempster, Laird, & Rubin, 1977). Details of this procedure and properties of the resulting chi-square statistic are described in Rudas, Clogg, and Lindsay (1994).

Examples

The first example is based on simulated data from a previous study (Zwick, Thayer, & Wingersky, 1994). The data consist of the item responses of 500 reference group (A) and 500 focal group (B) members. The reference group ability distribution was standard normal N(0, 1), while the focal group distribution was N(0.5, 1). The item responses were generated using a three-parameter logistic model (Birnbaum, 1968).

$$(5) \qquad P(\theta) = c + \frac{1 - c}{1 + \exp(-1.7a(\theta - b))},$$

where $P(\theta)$ is the probability of answering the item correctly for an examinee with ability θ. The item used in the example had a lower asymptote of c=0.15 and a discrimination of a=1 in both groups. The reference group difficulty was b_R=0 and the focal group difficulty was b_F=0.35. The item response functions for the reference and focal groups differed only in location; conditional on ability, the item was more difficult for the focal group. The measure of ability that served as a matching variable was the number-correct score on a 75-item test that included the example item.

For this analysis, the data can be summarized in a 76x2x2 contingency table. The sufficient statistics for π^* under (2) or (3) are the 76 (j=0,...,75) observed conditional odds ratios ($\hat{\alpha}_j$) and the frequencies g_{smj} and h_{smj}.

Out of the 304 observed frequencies, 103 were equal to zero; i.e. over one third of the cells were empty. Moreover, out of the 76 conditional 2x2 tables, 45 contained at least one zero frequency; therefore more than half of the 76 conditional odds ratios were impossible to estimate from the data or yielded estimated values of zero. Eliminating the 2x2 tables that contained empty cells would have required the deletion of 351 observations - over one third of the sample - which would not have been desirable.

To overcome the problem of empty cells we replaced the zero frequencies with small positive values. To assess the effect of this approach, the main analysis was carried out with various choices of the flattening (or smoothing) values. The values were either

constant (0.0001, 0.001, 0.01, 0.1, or 0.5), or uniformly distributed random on an interval starting at 0 and with the same expected values as above.

The estimates of π^* for the hypotheses of no DIF and uniform DIF, using the above flattening values, are reported in Table 1. The main finding is that, for every choice of the flattening values, $\hat{\pi}^*$(no DIF) and $\hat{\pi}^*$(uniform DIF) are very close to each other.

*** insert Table 1 around here ***

The numerical results in Table 1 show that increases in the flattening values result in decreases in the estimates for π^* (for the flattening values included). Estimates for π^* under both hypotheses have their minima near the flattening constant 0.9, where the estimates are 0.06064 and 0.06057, respectively. Taking into account, however, that several observed frequencies were equal to 0 or 1, it appears that 0.9 is too big to be used as a flattening constant.

The results in Table 1 show that we estimate that about 7% of the population needs to be disregarded in order to remove DIF, or about 93 % the population can be described by the model of no DIF. The actual choice of the flattening constant has very little effect on this result. Rudas, Clogg, and Lindsay (1994) described a method of obtaining lower confidence bounds for π^*, With this data set, using the flattening value of 0.1, one obtains the 95% lower confidence bound of 0.055 (rounded value) for π^*(no DIF). As the resulting 95% confidence interval does not contain zero, our procedure detects the DIF present in the original data generating mechanism.

The difference

(6) $\pi^*(\text{no DIF}) - \pi^*(\text{uniform DIF})$

can be used as a measure of the gain in fit due to using the model of uniform DIF over the model of no DIF. This quantity compares the estimates of the fractions of the population that cannot be described by the respective models. Although developing a formal test for the significance of this quantity is outside of the scope of the present paper, the results in Table 1 suggest that there is no substantial gain in using the model of uniform DIF to describe the data, compared to using the model of no DIF; in both cases we estimate that about 7% of the entire population (reference plus focal) cannot be described by the model.

In what follows, results using the flattening constant 0.1 will be described to illustrate the conclusions that can be reached using the π^* approach. The following table shows the 2x2 marginal of $\hat{\Psi}$ for the hypothesis of no DIF multiplied by the sample size. These are the observations that have to be removed in order to achieve conditional independence.

		Response	
		Correct	Incorrect
Group	Reference	7.85	12.57
	Focal	14.14	35.82

These may be compared with the corresponding marginal of the observed data:

		Response	
		Correct	Incorrect
Group	Reference	279	221
	Focal	326	174

This shows that we estimate over 20% (35.82/174) of focal group members who answered the item incorrectly to be outside the model of no DIF, while in the other categories, the fractions are much smaller. The observations that were removed from among focal group members who answered the item incorrectly account for more than 50% (35.82/70.38) of the total number of observations that must be removed. This means that although the model of uniform DIF does not describe our data substantially better than the model of no DIF, the model of no DIF fails to account for some focal group members who did not answer correctly. This indicates the presence of some degree of DIF in favor of the reference group, as in the original mechanism of data generation. Note that the estimate of Ψ under the hypothesis of uniform DIF is very similar to the estimate under the hypothesis of no DIF and has the same interpretation as above. That is, both the magnitude of the misfit (as measured by π^*) and the pattern of residuals are similar for the two hypotheses.

Under the hypothesis of uniform DIF, the value of the conditional odds ratio for which the minimum occurred is $\hat{\alpha}$(uniform DIF)=1.09375. There are only two types of conditional tables in which the pattern of decreases in cell counts is different for the no DIF and uniform DIF hypotheses: (1) tables in which one of the hypotheses holds exactly and (2) tables in which $\hat{\alpha}_j$ is between α=1 and $\hat{\alpha}$(uniform

DIF).

The value of $\hat{\alpha}$(uniform DIF) is equal to the odds ratio that was observed among those who had 44 correct answers. An interesting interpretation of this value can be obtained by noting that, out of the 1000 observations, 473 were in conditional tables where the estimated conditional odds ratio (after replacing each zero by 0.1) was less than 1.09375, 24 were in the conditional table where the estimated conditional odds ratio was exactly 1.09375 and 503 came from tables where the estimated conditional odds ratio was greater than 1.09375. This means that the π^* approach led to a median-type estimate of the common conditional odds ratio.

Plotting $\hat{\Psi}$ against the number correct score may be informative in revealing the pattern of occurrence of DIF, but, because of the small value of π^*, we did not apply this technique here. Note that for examinees with at least 47 correct answers, only the frequencies of the cells with incorrect responses were reduced (under both hypotheses).

The conventional MH DIF analysis involves calculation of the MH chi-square and the index

$$\text{MH D-DIF} = -2.35(\ln\hat{\alpha}_{MH}),$$

a transformation of the MH odds ratio estimate, $\hat{\alpha}_{MH}$, to the delta metric of item difficulty (Holland, & Thayer, 1988).

For the (unsmoothed) example data, the MH chi-square statistic is 0.30, $\hat{\alpha}_{MH}$=1.11, and MH D-DIF is -0.24, with a standard error of 0.38 (see Phillips & Holland, 1987). Since the chi-square statistic is

close to zero and MH D-DIF is close to its null value of zero, the conclusion from the MH analysis is that there is no reason to reject the hypothesis of no DIF. That is, the MH method fails to detect the DIF in the population, in contrast with the π^* approach.

The data for the second example were taken from the 1993 Advanced Placement Physics B Exam. There were 70 multiple choice items and the goal of the analysis was to detect male/female DIF. There were data available on 9104 male (reference group) and 4118 female (focal group) examinees. The matching variable was the number-correct score on the 70 items. Only results for the first 10 items will be reported here. Zero observed frequencies were replaced by 0.1, as in the previous analysis.

*** Insert Table 2 around here***

The results are summarized in Table 2. For the 10 items considered, the $\hat{\pi}^*$ values for the no-DIF hypothesis are between 0.02 and 0.06, and for the uniform-DIF hypothesis between 0.02 and 0.04, i.e. we estimate that for each item, DIF is absent in 94-98% of the population, and uniform DIF characterizes 97-98% of the population. The values of (6), showing the gain in fit due to assuming uniform DIF instead of no DIF, are between 0.00 and 0.03. For items 1, 2, 5, 9, and 10, the uniform-DIF hypothesis does not fit better, as measured by the π^* index of fit, than the no-DIF hypothesis. The gain is the highest for items 3, and 7, namely 3%. Whether this gain should be considered substantial or not, may depend on several factors. One possible approach is to consider the ratio $\hat{\pi}$(uniform DIF)/$\hat{\pi}$(no DIF). This shows that for items 3, and 7, the fraction of the population not described reduced by 50% as one moves from the no-DIF hypothesis to the uniform-DIF hypothesis.

Except for items 3, 8, and 10, the $\hat{\alpha}$(uniform DIF) values suggest superior item performance for males conditional on number-correct score. The magnitude of DIF is greatest (above 2) for item 4. Assuming a uniform DIF of this magnitude, leads to the description of an estimated 98% of the total population. No other assumed value of the common conditional odds ratio could lead to the description of a greater fraction of the population.

There are several further analyses that are facilitated by the π^* approach. For example, in the case of item 4, DIF appears to be concentrated at lower ability levels, and, consequently, examinees at higher ability levels are affected by DIF to a lesser degree. It was found, that 81% of the individuals who could not be described by the no-DIF hypothesis had number-correct scores below the median. Ninety-five percent of those who could not be described by the no-DIF hypothesis had number correct-scores below the 75th percentile. The corresponding figures for item 10 are 79% and 91% respectively, showing again a concentration of DIF at lower ability levels. All 10 items showed the same effect to some degree.

*** Insert Table 3 around here ***

Results of the MH analysis are reported in table 3. Items 3 and 8 had odds ratios less than one, indicating that females tended to perform better, conditional on number-correct score, whereas the other items showed better conditional item performance for males. Using ETS criteria (Zieky, 1993), only item 4 shows substantial DIF against females.

The analyses based on the π^* approach and on the MH method agree

considerably as to the estimates of the common conditional odds ratios for all the 10 items of the test considered. In the case of item 10, the two analyses disagree concerning the direction of DIF; However, the estimated common conditional odds ratios are close to one in both analyses, and in the MH approach the result is not significant. However, the strength or importance of DIF is conceptualized in very different ways in the two approaches: the magnitude and statistical significance of the odds ratio estimate in the MH analysis versus the size of the fraction of the population that cannot be described by the hypothesis of interest in the π^* approach.

Discussion

The π^* approach offers a new way to assess the importance of DIF in educational testing. The importance of DIF, in this approach, is influenced by the size of the subgroup of the population in which DIF may be present, as well as the magnitude of DIF for this subpopulation. In this sense, the results of the π^* method, when applied to the problem of DIF, will depend to some degree on the distribution of the observations in the reference and focal groups, and the distribution of the matching variable. Note that the MH odds ratios are also affected by the distribution of the examinees. The MH odds ratio estimate can be expressed as a weighted sum of the $\hat{\alpha}_j$ values, where the weights are a function of the observed within-level cell frequencies (Holland & Thayer, 1988). In addition, the examinee ability distribution can have unintended effects on the MH odds ratios (Zwick, 1990).

The π^* approach gives results with a straightforward interpretation,

and may provide diagnostic information concerning the specific parts of the population where DIF is evident. By inspecting $\hat{\Psi}$, it might be found, for example, that the lack of fit of the no-DIF hypothesis tended to occur among examinees who chose a particular incorrect response. This type of information could be helpful in pinpointing the source of DIF. Or it might be found that lack of fit to the no-DIF hypothesis occurred only among examinees in the extremes of the test score distribution. This might be viewed as less consequential than DIF occurring near the mean of the distribution.

Finally, it should be mentioned that recent research (Chang, Mazzeo, & Roussos, 1995; Roussos & Stout, 1993) has shown that under some circumstances, the SIBTEST method of DIF detection (Shealy, & Stout, 1993) maintains better Type I error control than MH-type methods. The π^* approach has no intrinsic connection to the MH method and could be applied in conjunction with SIBTEST or with other DIF detection methods as well.

References

Agresti, A. (1990) Categorical Data Analysis. Wiley.

Birnbaum, A. (1968) Some latent trait models. In F. Lord., M.Novick (eds.) Statistical theories of mental test scores. Reading: Addison-Wesley.

Camilli, G., & Shepard, L. A. (1994) Methods for Identifying Biased Test Items. Thousand Oaks: Sage.

Chang, H.-H., Mazzeo, J., & Roussos, L. (1995) Detecting DIF for

polytomously scored items: An adaptation of the SIBTEST procedure. ETS Research Report 95-5. Princeton, NJ: ETS.

Clogg, C. C. (1981) Latent structure models of mobility. American Journal of Sociology, 86, 836-868.

Clogg, C. C., Rudas, T., & Xi, L. (1995) A new index of structure for the analysis of models for mobility tables and other cross-classifications. To appear in Marsden, P. (ed.) Sociological Methodology 1995.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society (Ser B), 39, 1-22.

Donoghue, J. R., & Allen, N. L. (1993) Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. Journal of Educational Statistics, 18, 131-154.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993) A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. in: P. W. Holland and H. Wainer (eds.) Differential item functioning, pp 137-166. Hillsdale, NJ: Erlbaum.

Goodman, L. A. (1979) Simple models for the analysis of cross-classifications having ordered categories. Journal of the American Statistical Association, 74, 537-552.

Holland, P. W. (1985) On the study of differential item performance without IRT. Proceedings of the Military Testing Association, October.

Holland, P. W., & Thayer, D. T. (1988) Differential item performance and the Mantel-Haenszel procedure. in: H. Wainer, H. I. Braun (eds.) Test Validity. Hillsdale, NJ: Erlbaum.

Holland, P. W., & Wainer, H. (eds.) (1993) Differential Item Functioning. Hillsdale, NJ: Erlbaum.

Mantel, N., & Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Mellenbergh, G. J. (1982) Contingency table models for assessing item bias. Journal of Educational Statistics, 7, 105-118.

Phillips, A., & Holland, P. W. (1987) Estimation of the variance of the Mantel-Haenszel log-odds-ratio estimate. Biometrics, 43, 425-431.

Roussos, L. A., & Stout, W. F. (1993, April) Simulation studies of effects of small sample size and studied item parameters on SIBTEST and Mantel Haenszel type I error performance. Presented at the annual meeting of the American Educational Research Association, Atlanta.

Rudas, T. (1991) Prescribed conditional interaction structure models with application to the analysis of mobility tables. Quality and Quantity, 25, 345-358.

Rudas, T, Clogg, C. C., & Lindsay, B. G. (1994) A new index of fit based on mixture methods for the analysis of contingency tables.

Journal of the Royal Statistical Society (Ser B), 54, 623-640.

Rudas, T., & Leimer, G.-H. (1992) Analysis of contingency tables with known conditional odds ratios or known log-linear parameters. In P. G. M. van der Heijden, W. Jansen, B. Francis, G. U. H. Seeber (eds.) Statistical Modelling. Amsterdam: North-Holland.

Shealy, R., & Stout, W. (1993) A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. Psychometrika, 58, 159-194.

Zieky, M. (1993) Practical questions in the use of DIF statistics in test development. In P. W. Holland and H. Wainer (eds.) Differential item functioning, pp. 337-347. Hillsdale, NJ: Erlbaum.

Zwick, R. (1990) When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? Journal of Educational Statistics, 15, 185-197.

Zwick, R., Donoghue, J. R., & Grima, A. (1993a) Assessment of differential item functioning for performance tasks. Journal of Educational Measurement, 30, 233-251.

Zwick, R., Donoghue, J. R., & Grima, A (1993b) Assessing differential item functioning in performance tests. ETS Research Report 1993-14.

Zwick, R., & Thayer, D. T. (in press) Evaluating the magnitude of differential item functioning in polytomous items. Journal of Educational and Behavioral Statistics.

Zwick, R., Thayer, D. T., & Wingersky, M. (1994) A simulation study of the methods for assessing differential item functioning in computerized adaptive tests. Applied Psychological Measurement, 18, 121-140.

Table 1

Maximum likelihood estimates of π^* for the hypotheses of no DIF and uniform DIF using different flattening values for the data generated by (5)

Flattening value	$\hat{\pi}^*$(no DIF)	$\hat{\pi}^*$(uniform DIF)
0.0001	0.07338	0.07274
0.001	0.07335	0.07272
0.01	0.07309	0.07243
0.1	0.07039	0.06953
0.5	0.06387	0.06339
U(0, 0.0002)	0.07339	0.07275
U(0, 0.002)	0.07338	0.07274
U(0, 0.02)	0.07332	0.07265
U(0, 1)	0.07071	0.06928

Table 2

Maximum likelihood estimates of

π^*(no DIF), π^*(uniform DIF), α(uniform DIF)

for the first 10 items of the 1993 Advanced Placement Physics B Exam

Item No	$\hat{\pi}^*$(no DIF)	$\hat{\pi}^*$(uniform DIF)	$\hat{\alpha}$(uniform DIF)
1	0.03	0.03	1.03
2	0.03	0.03	1.08
3	0.05	0.02	0.63
4	0.04	0.02	2.08
5	0.03	0.03	1.24
6	0.03	0.02	1.28
7	0.06	0.03	1.62
8	0.04	0.03	0.87
9	0.02	0.02	1.15
10	0.02	0.02	0.92

Table 3

Results of the Mantel-Haenszel analysis for the first 10 items of the 1993 Advanced Placement Physics B Exam

Item No	MH odds ratio	MH D-DIF	standard error of MH D-DIF
1	1.07	-0.16	0.11
2	1.10	-0.23	0.10
3	0.69	0.89	0.11
4	2.00	-1.63	0.13
5	1.02	-0.05	0.10
6	1.16	-0.35	0.10
7	1.49	-0.94	0.10
8	0.87	0.32	0.10
9	1.26	-0.54	0.12
10	1.08	-0.19	0.12